Unsupervised Discriminant Projection Analysis for Feature Extraction

Jian Yang David Zhang
Biometrics Centre, Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong
{csjyang,csdzhang}@comp.polyu.edu.hk

Zhong Jin Jing-yu Yang
Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, P. R. China
yangjy@mail.njust.edu.cn

Abstract

This paper develops an unsupervised discriminant projection (UDP) technique for feature extraction. UDP takes the local and non-local information into account, seeking to find a projection that maximizes the non-local scatter and minimizes the local scatter simultaneously. This characteristic makes UDP more intuitive and more powerful than the up-to-date method — Locality preserving projection (LPP, which considers the local information only) for classification tasks. The proposed method is applied to face biometrics and examined using the ORL and FERET face image databases. Our experimental results show that UDP consistently outperforms LPP, PCA, and LDA.

1. Introduction

PCA and LDA are two well-known linear subspace learning techniques and have become the most popular methods for face recognition [1-3]. Recently, He et al [5, 6] proposed a method called Locality Preserving Projections (LPP) and applied it to face recognition. LPP is a linear subspace method derived from Laplacian Eigenmap [4]. It results in a linear map that optimally preserves local neighborhood information in a certain sense. In contrast to most manifold learning algorithms, a remarkable advantage of LPP is that it can generate a simple and efficiently-computable linear map, like that of PCA or LDA.

LPP is modeled based on the characterization of "locality". This modeling, however, has no direct connection to classification. The objective function of LPP is to minimize the local quantity, i.e., the local scatter of the projected data. This criterion cannot guarantee to yield a good projection for classification in some cases where the "non-locality" provides dominant information for discrimination. In this paper, we will address this problem and explore a more effective projection for classification purpose. We will consider two quantities, local and non-local, at the same time in the modeling process.

We first present the techniques to characterize the local and non-local scatters of data. Then, based on this characterization, we propose a criterion, which seeks to maximize the ratio of the non-local scatter to the local scatter. This criterion, similar to the classical Fisher criterion, is a Rayleigh quotient in form. Thus, it is not hard to find its optimal solutions by solving a generalized eigen-equation. Since the proposed method does not use the class-label information of samples in the learning process, this method is called unsupervised discriminant projection (UDP), in contrast with the supervised discriminant projection of LDA.

In contrast with LPP, UDP has intuitive relations to classification since it utilizes the information of the "non-locality". Provided that each cluster of samples in the observation space is exactly within a local neighbor, UDP can yield an optimal projection for clustering in the projected space, while LPP cannot. As a feature extraction method, UDP will be demonstrated more effective than LPP, PCA and LDA, based on our experiments using two face image databases.

2. Unsupervised Discriminant Projection (UDP) Analysis

2.1. Characterization of the Local Scatter

Recall that in PCA, in order to preserve the global geometric structure of data in a transformed low-dimensional space, the global scatter of samples is considered. Instead, if we aim to discover the local structure of data, the local scatter (or intra-locality scatter) of samples should be considered. The local scatter can be characterized by the mean square of the Euclidean distance between any pair of the projected sample points that are within any local δ -

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

neighborhood ($\delta > 0$). Specifically, two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are viewed within a local δ -neighborhood provided that $\| \mathbf{x}_i - \mathbf{x}_j \|^2 < \delta$. Let us denote the set $\mathrm{U}^{\delta} = \{(i,j) \big| \| \mathbf{x}_i - \mathbf{x}_j \|^2 < \delta\}$. After the projection of $\mathbf{x}_i$ and $\mathbf{x}_j$ onto a direction $\mathbf{w}$, we get their images y_i and y_j . The local scatter of is then defined by

$$J_L(\mathbf{w}) \triangleq \frac{1}{2}\frac{1}{M_L}\sum_{(i,j)\in \mathrm{U}^{\delta}}(y_i - y_j)^2 \propto \frac{1}{2}\sum_{(i,j)\in \mathrm{U}^{\delta}}(y_i - y_j)^2 \quad (1)$$

where M_L is the number of sample pairs satisfying $\| \mathbf{x}_i - \mathbf{x}_j \|^2 < \delta$.

Let us define the adjacency matrix **H**, whose element is given below:

$$H_{ij} = \begin{cases} 1, & \| \mathbf{x}_i - \mathbf{x}_j \|^2 < \delta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

It is obvious that the adjacency matrix **H** is a symmetric matrix.

By virtue of the adjacency matrix **H**, Eq. (1) can be rewritten by

$$J_L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}(y_i - y_j)^2 \quad (3)$$

It follows from Eq. (3) that

$$J_L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2 = \mathbf{w}^T\mathbf{S}_L\mathbf{w}, \quad (4)$$

where

$$\mathbf{S}_L = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \quad (5)$$

$\mathbf{S}_L$ is called the local scatter (covariance) matrix.

Due to the symmetry of **H**, it follows that

$$\mathbf{S}_L = \frac{1}{2}\left(\sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}\mathbf{x}_i\mathbf{x}_i^T + \sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}\mathbf{x}_j\mathbf{x}_j^T - 2\sum_{i=1}^{M}\sum_{j=1}^{M} H_{ij}\mathbf{x}_i\mathbf{x}_j^T\right) = (\mathbf{X}\mathbf{D}\mathbf{X}^T - \mathbf{X}\mathbf{H}\mathbf{X}^T) = \mathbf{X}\mathbf{L}\mathbf{X}^T, \quad (6)$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M)$, and **D** is a diagonal matrix whose elements on diagonal are column (or row since **H** is a symmetric matrix) sum of **H**. $\mathbf{L} = \mathbf{D} - \mathbf{H}$ is called Laplacian matrix in [4-6].

It is obvious that **L** and $\mathbf{S}_L$ are both real symmetric matrices. From Eqs. (4) and (6), we know $\mathbf{w}^T\mathbf{S}_L\mathbf{w} \geq 0$ for any nonzero vector $\mathbf{w}$. So, the local scatter matrix $\mathbf{S}_L$ must be non-negative definite.

In the above discussion, we use δ -neighborhoods to characterize the "locality" and the local scatter. This way is geometrically intuitive but unpopular because it is hard to choose a proper neighborhood radius δ in practice. To void the difficulty, the method of K-nearest neighbors is always used instead in real-world applications. The K-nearest neighbors method can determine the following adjacency matrix **H**, with elements given by:

$$H_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \text{ is among K nearest nieghbors of } \mathbf{x}_i \\ & \textit{and } \mathbf{x}_i \text{ is among K nearest nieghbors of } \mathbf{x}_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

The local scatter can be characterized similarly by K-nearest neighbor adjacency matrix if Eq. (2) is replaced by Eq. (7).

2.2. Characterization of the Non-local Scatter

In contrast to the characterization of the local scatter, the non-local scatter (i.e., the inter-locality scatter) can be characterized by the mean square of the Euclidean distance between any pair of the projected sample points that are outside any local δ -neighborhood ($\delta > 0$).

Let us denote the set $\mathrm{U}_N^{\delta} = \{(i,j) \big| \| \mathbf{x}_i - \mathbf{x}_j \|^2 \geq \delta\}$. The non-local scatter is defined by

$$J_N(\mathbf{w}) \triangleq \frac{1}{2}\frac{1}{M_N}\sum_{(i,j)\in \mathrm{U}_N^{\delta}}(y_i - y_j)^2 \propto \frac{1}{2}\sum_{(i,j)\in \mathrm{U}_N^{\delta}}(y_i - y_j)^2 \quad (8)$$

where M_N is the number of elements in U_N^{δ}.

By virtue of the adjacency matrix **H** in Eq. (2) or (7), the non-local scatter can be rewritten by

$$J_N(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(1 - H_{ij})(y_i - y_j)^2 \quad (9)$$

It follows from Eq. (9) that

$$J_N(\mathbf{w}) = \mathbf{w}^T\left[\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(1 - H_{ij})(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\right]\mathbf{w} = \mathbf{w}^T\mathbf{S}_N\mathbf{w}, \quad (10)$$

where

$$\mathbf{S}_N = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(1 - H_{ij})(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \quad (11)$$

$\mathbf{S}_N$ is called the non-local scatter (covariance) matrix. It is easy to show $\mathbf{S}_N$ is also a non-negative definite matrix.

Let us define the matrix $\mathbf{H}_N = (1 - H_{ij})_{M\times M}$. Similar to the derivation of Eq. (6), we have

$$\mathbf{S}_N = \mathbf{X}\mathbf{L}_N\mathbf{X}^T \quad (12)$$

where $\mathbf{L}_N = \mathbf{D}_N - \mathbf{H}_N$, $\mathbf{D}_N$ is a diagonal matrix whose elements on diagonal are column (or row) sum of $\mathbf{H}_N$.

2.3. Criterion of UDP

The technique of Locality Preserving Projection (LPP) [5] seeks to find a linear subspace which can preserve the local structure of data. LPP is actually to minimize the local scatter $J_L(\mathbf{w})$. Intuitively, the projection direction determined by LPP can ensure that, if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are close, their projections y_i and y_j are close as well. But, LPP cannot guarantee that, if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are not close, their projections y_i and y_j are not either. This means, it possibly happens that two faraway samples belonging to different classes may result in close images after the projection of LPP. Therefore, LPP does not necessarily yield a good projection suitable for classification.

For the purpose of classification, an intuitive motivation is to find a projection, which makes the close samples become closer and simultaneously make the distant samples become more distant. From this point of view, a desirable projection should be the one that minimizes the local scatter $J_L(\mathbf{w})$ and maximizes the non-local scatter $J_N(\mathbf{w})$ at the same time. Actually, we can obtain such a projection by maximizing the following criterion:

$$J(\mathbf{w}) = \frac{J_N(\mathbf{w})}{J_L(\mathbf{w})} = \frac{\mathbf{w}^T\mathbf{S}_N\mathbf{w}}{\mathbf{w}^T\mathbf{S}_L\mathbf{w}} \quad (13)$$

The criterion in Eq. (13) is formally similar to the Fisher criterion since they are both Rayleigh quotients. Differently, the matrices $\mathbf{S}_L$ and $\mathbf{S}_N$ in Eq. (13) can be constructed without knowing the class-label of samples while the between-class and within-class scatter matrices in the Fisher criterion cannot. This means the Fisher discriminant projection is supervised while the projection determined by $J(\mathbf{w})$ can be obtained in an unsupervised manner. So, this projection is called Unsupervised Discriminant Projection (UDP) in this paper.

2.4. Algorithm of UDP

If the local scatter matrix $\mathbf{S}_L$ is non-singular, the criterion in Eq. (13) can be maximized directly by calculating the generalized eigenvectors of the following generalized eigen-equation:

$$\mathbf{S}_N\mathbf{w} = \lambda\mathbf{S}_L\mathbf{w} \quad (14)$$

The projection axes of UDP can be selected as the generalized eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d$ of $\mathbf{S}_N\mathbf{w} = \lambda\mathbf{S}_L\mathbf{w}$ corresponding to d largest positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

After obtaining the projection axes, we can form the following linear transform for a given sample $\mathbf{x}$:

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}, \text{ where } \mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d) \quad (15)$$

The feature vector $\mathbf{y}$ is used to represent the sample $\mathbf{x}$ for recognition purpose.

In real-world biometrics applications such face recognition, however, $\mathbf{S}_L$ is always singular due to the given limited amount of training samples. In such cases, the classical algorithm cannot be used directly to solve the generalized eigen-equation. To avoid this difficulty, we can adopt the two-phase strategy used in Fisherfaces or Laplacianfaces. That is, PCA is first used for dimension reduction and then UDP is performed in the PCA-transformed space.

3. Experiments

3.1. Experiment Using the ORL Database

The ORL (or called AT&T) database contains face images from 40 subjects, each providing 10 different images. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details. The size of each image is 92x112 pixels, with 256 grey levels per pixel.

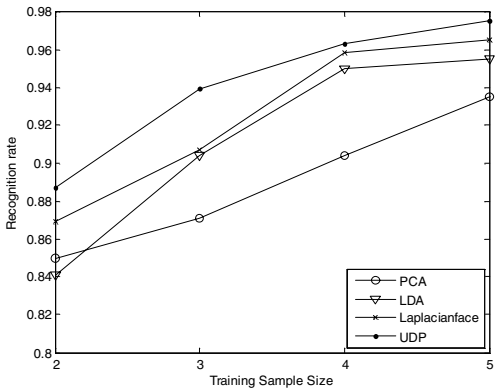

Figure 1. The maximal average recognition rates of four methods versus the variation of the training sample sizes

In our experiments, the first *l* images (*l* varies from 2 to 5) of each individual are used for training, and the remaining (10 − *l*) images are used for test. For each *l*, PCA (Eigenface) [1], LDA (Fisherface) [3], LPP (Laplacianface) [6] and the proposed UDP are, respectively, used for feature extraction. In the PCA phase of LDA, Laplacianface and UDP, the number of principal components is set as 25, 40, 50, and 60, respectively corresponding to *l* = 2, 3, 4, and 5. The K-nearest neighborhood parameter K in Laplacianface

and UDP is chosen as K = $l - 1$. Finally, a nearest-neighbor classifier with cosine distance is employed for classification. The recognition rate curve versus the variation of training sample sizes is shown in Figure 1. Figure 1 indicates UDP consistently performs better than Laplacianface, LDA, and PCA as the training sample size varies from 2 to 5. When the training sample size l = 5, the recognition rate of UDP is up to 97.5%. This result is very encouraging in contrast to the previous ones on this database.

3.2. Experiment Using the FERET Database

The final experiment is performed on a subset of the FERET database [7-9], which includes 1000 images of 200 individuals (each one has 5 images). It is composed of the images whose names are marked with two-character strings: "*ba*", "*bj*", "*bk*", "*be*", "*bf*". This subset involves variations in facial expression, illumination, and pose. In our experiment, the facial portion of each original image was automatically cropped based on the location of eyes and mouth, and the cropped image was resized to 80×80 pixels and pre-processed by histogram equalization.

In our test, we use the first two images (i.e., "ba" and "bj") per class for training, and the remaining three images (i.e., "bk", "be" and "bf") for test. PCA, LDA, Laplacianface and UDP are, respectively, used for feature extraction. In the PCA phase of LDA, Laplacianface and UDP, the number of principal components is set as 120. The K-nearest neighborhood parameter K in Laplacianface and UDP is chosen as K = $l - 1 = 1$. After feature extraction, a nearest neighbor classifier with cosine distance is employed for classification. The maximal recognition rate of each method and the corresponding dimension are given in Table 1. Table 1 demonstrates again that UDP outperforms PCA, LDA and Laplacianface.

Table 1. The maximal recognition rates (%) of the four methods on a subset of FERET database and the corresponding dimensions

Method	PCA	LDA	LPP	UDP
Accuracy	73.3	75.0	77.0	80.5
Dimension	85	100	105	100

4. Conclusions

We develop an unsupervised discriminant projection (UDP) technique for feature extraction in this paper. UDP takes account of the local and non-local scatters at the same time and seeks to find a projection maximizing the ratio of the non-local scatter to the local scatter. The utilization of the non-local information makes UDP more intuitive and more powerful than LPP for classification tasks. Our experimental results on two face image databases demonstrate that UDP is more effective than LPP (Laplacianface), PCA and LDA.

Acknowledgements

This research was supported by the National Science Foundation of China under Grants No. 60503026, No. 60332010, No. 60472060, and No. 60473039, and the CERG fund from the HKSAR Government and the central fund from the Hong Kong Polytechnic University.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition", *J. Cognitive Neuroscience*, 1991, 3(1), pp.71-86.

[2] D. L. Swets and J. Weng. "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. Pattern Anal. Machine Intell.*, 1996,18(8), pp. 831-836.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection", *IEEE Trans. Pattern Anal. Machine Intell.* 1997, 19 (7), pp. 711-720.

[4] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", *Neural Computation*, 2003, 15 (6), 1373-1396.

[5] X. He, P. Niyogi, "Locality Preserving Projections", *Neural Information Processing Systems 16* (NIPS'2003).

[6] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, "Face Recognition Using Laplacianfaces", *IEEE Trans. Pattern Anal. Machine Intell.*, 2005, 27(3), pp. 328-340.

[7] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms", *IEEE Trans. Pattern Anal. Machine Intell.*, 2000, 22 (10), pp.1090-1104.

[8] P. J. Phillips, "The Facial Recognition Technology (FERET) Database", http://www.itl.nist.gov/iad/humanid/feret/feret_master.html

[9] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing* 1998, 16 (5), pp. 295-306.